AD_______________

Award Number: W81XWH-12-C-0236

TITLE: Electrical Impedance Imaging for Early Detection of Breast Cancer for Young Women

PRINCIPAL INVESTIGATOR: Alexander Stojadinovic

CONTRACTING ORGANIZATION: The Henry M Jackson Foundation for the Advancement of Military Medicine, Inc
Bethesda, MD. 20817

REPORT DATE: December 2013

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE
December 2013

2. REPORT TYPE
Final

3. DATES COVERED
27September2012-26September2013

4. TITLE AND SUBTITLE
Electrical Impedance Imaging for Early Detection of Breast Cancer for Young Women

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-12-C-0236

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Alexander Stojadinovic

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Henry M Jackson Foundation for the Advancement of Military Medicine, Inc
Bethesda, MD. 20817

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The Electrical Impedance Scanning (EIS) Breast Cancer Study focuses on evaluating a new technology, electrical impedance scanning (EIS), as a tool for identifying women; who, because of breast tissue changes, are also at elevated risk of developing breast cancer, and who, thereby, would benefit from increased surveillance and the initiation of early breast imaging. The specific aim of this effort is to develop a finally Bayesian Belief Network (BBN) classification model to estimate the interval risk of malignancy and pre-malignancy in young women. A selected wound outcomes dataset, including cytokine and chemokine data will be analyzed to construct a series of predictive models. These models were cross-validated for statistical significance and interpreted for clinical significance.

15. SUBJECT TERMS- Bayesian Belief Network (BBN); electrical impedance scanning; breast cancer, breast density, palpable lesion, nulliparity, history (current or past) of Depo use, history of hormone replacement therapy, prior biopsy, frequency of menstrual periods

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			USAMRMC
U	U	U	UU		**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

Introduction

The Electrical Impedance Scanning (EIS) Breast Cancer Study focuses on evaluating a new technology, electrical impedance scanning (EIS), as a tool for identifying women, who, because of breast tissue changes, are also at elevated risk of developing breast cancer, and who, thereby, would benefit from increased surveillance and the initiation of early breast imaging. The specific aim of this effort is to develop a final Bayesian Belief Network (BBN) classification model to estimate the interval risk of malignancy and pre-malignancy in young women. A selected wound outcomes dataset, including cytokine and chemokine data was analyzed to construct a series of predictive models. These models were then cross-validated for statistical significance and interpreted for clinical significance.

Body

Development of a Master Subject Index

The original database was de-identified and the updated consent forms contained only personally identifiable information. To facilitate the matching of the updated consents to the correct records in the database a master subject index (MSI) was created. The MSI was based on a list that was received with some personally identifiable information such as first and last names and the addresses of each patient in the study, the date of initial consent, and the record ID number. The physical re-consent forms were then reviewed and the patient's updated consent status was added.

As manual remediation commenced it became clear that the MSI as described above was not sufficient to properly match the updated consent status back to the corresponding records in the database. The record ID numbers in the MSI referred to the physical file ID and not the database record ID. Although in many cases the two numbers were the same, there was a large number of files where an exact ID match did not exist. Notably the number of subjects in the MSI and the database differed extensively, with the database consisting of 4033 records while the MSI had 4691 records. As there were more records in the MSI than there were in the database it was decided that it would be more efficient to work from the list of unmatched database IDs to try and find the missing matches.

The majority of the database IDs were of the form SITE#### where # represents a numeric character and SITE represents the long form abbreviation of the study site (e.g. WRAMC1094). 806 database IDs did not follow this convention. In these cases, the IDs are four digit numbers that range from 1647 to 2474 (almost completely sequential) and there is no notable pattern in how these numbers were assigned (i.e. the order of numbers is not based on study site or date of first consent). Looking at the pattern of box IDs, some of the IDs often contained site abbreviations other than those used in the database (i.e. wr instead of WRAMC, wp instead of KACH) and the length of the numeric part of the ID often consisted either of a single 1-4 digit number or two 1-4 digit numbers separated by a dash.

Based on the insight of box ID format a computer function was written that for all unmatched IDs in the MSI (box IDs) Database IDs were created by combining the proper abbreviation of the study site and a the numeric part of the box ID with leading zeros added until the number of digits reached four. For example, the suggested database ID for a box ID MG271 would be MGAFMC0271. For records where the numeric part of the MSI ID consisted of two numbers separated by a dash, the latter number was used after confirming this was the case with at least one record. A separate piece of code was written that took all unmatched IDs from the database and compared them to the newly generated guesses. If the MSI guess ID matched that database ID exactly and the date of the first visit in the database matched the date of first consent in the MSI exactly, it was considered to be a match. For those IDs where the database ID matched the guess MSI ID exactly but the dates did not match, physical files were used to confirm whether or not the information in the file matched the information in the database for a few key fields (e.g. age at enrollment, race, last four digits of social security number). If it did, that match was recorded.

For the database IDs that were still unmatched at this point (1,370 records) an algorithm was created to come up with additional guesses from the unmatched ID by looking for all MSI IDs where the study site matched and the date of first visit in the database deviated by no more than a month from the date of first consent in the MSI. To make the search more efficient the list was transformed to include every MSI (Box) ID and then as guesses listed all database IDs where the algorithm returned the MSI as a possible ID match. Physical files were examined to match the information for selected key fields with the information in the database to identify the correct guess.

Following the above process, there were still 502 database IDs where the correct box ID match had not been identified. A running list was maintained of these IDs and during remediation whenever a file was found where the box ID could not be located in the database, a database ID that matched the information in the file was determined. Using this process, all files were successfully matched to a database record and a box ID was found for all but 16 database records (all of which contained no information).

This allowed the creation of a final de-identified Master Subject Index that listed the box ID, database ID, date of first visit, date of first consent, type of updated consent and cut-off date for what data could be used for that patient.

Audits of the study database

Before fully initiating data remediation three audits of the study database were performed; one electronic audit and two manual audits were conducted.

The electronic audit was conducted to get a schematic overview of what was represented in the study database and to identify some of the problems that needed to be addressed. Since no data dictionary existed for the study database one of the main goals of the electronic audit was to create one. For each of the tables in the study database the distribution of each variable was examined; identified issues such as outliers, lack of standardization, and unusual values; ideas on how to treat the variable moving forward and when possible a definition for the variable was developed.

The first manual remediation focused on both gaining insights of the formatic relationship between the data in the physical patient files and the data in the database and performing a preliminary analysis on how accurately the database represented the information from the physical file data. The location of the information in the physical files was identified in order to find the information needed to fill in each variable and to determine how consistent file formatting was across all records. To accomplish this, 100 patients were randomly selected and identified and it was recorded where in the file the data could be found for any given variable and was then compared the values of the file to the values in the database, recording any recorded any discrepancies. In retrospect the sample size for this audit should have likely been

somewhat bigger as the results of this audit did not accurately represent the inconsistencies in file format and completion both between the different study sites and within each site.
Before conducting the second electronic remediation results of both of the aforementioned audits were used, as well as results of the DCI audit in order to form a remediation plan. For every variable in the database a reasoned decision was made of whether or not the variable should be included in the updated database and if so what type of remediation did the variable require. Variables were considered for elimination if they met any of the following criteria: due to change in study objective (i.e. the switch from EIS functionality to risk based screening stratification) the variable was no longer relevant, the variable was a duplicate, the variable was empty. The variables not selected for removal were then subject to the second audit so that type and extent of remediation could be better decided.

During the second audit 100 patients were random selected, making sure that the study site proportions in the sample mirrored those in the database. For each patient data was re-entered into the file regardless of whether the database value was incorrect for the time period acceptable according to the type of consent (one year from date of initial consent for non-re-consented patients and everything for re-consented patients). This audit revealed many inconsistencies between visit records in the database and the actual files in that there were several visit records in the files that were not in the database and several visits in the database that were not in the actual files. Error rates were then calculated for all variables and based on those results updated the remediation plan. Both variables were accounted for where the error rate exceeded what had been estimated and needed to be flagged for manual remediation and variables that were based on previous information were erroneously flagged for manual remediation when it proved to be not necessary. After the second manual audit was completed we created an updated data dictionary and specs for re-entry in the database.

Manual Data Remediation

Tables were imported from the original database into a Microsoft Access database and were reformatted according to the remediation plan, new ´correction´ variables were added and forms were created for both correcting data and for entering new records.

The manual remediation consisted of carefully reading each physical file and comparing its contents to that of the database. For patients that were re-consented the entire file was examined for accuracy. For patients that only had initial consent all dates were first validated then all records that fell within a year of enrollment were reviewed. All records for patients whose consent had been withdrawn were marked for removal. When a value for a variable in the database did not match the value in file the accurate value from the file was entered into the corresponding correction variable in the database. Any record (visit, exam, surgery, etc.) that was in the database but could not be found in the files was marked for removal and any record that was in the files but not in the database was entered into the database through forms containing 'correction' variables not only for the variables being manually remediated but for all variables not to be deleted.

At one point here were two people concurrently doing manual remediation and there were two versions of the manual remediation database that were later combined once the process was completed.

Electronic Data Remediation, Data Curation and Flattening

Figure 1 in **Appendix A** shows a concentrated graphical depiction of the data curation and electronic remediation process. The remainder of this section consists of a more detailed description of this process.
To maximize time and cost efficiency the remediation was planned to minimize what had to be manually fixed to only those variables where it was absolutely needed. This meant that electronic remediation and data curation had to compensate both with simplification, and cleaning and varying levels of combination to get data from all sources into a single database. Since two separate manual remediation databases were being used (one of which was at one point a copy of the other with some overlap of corrected values), there was a need to somehow merge them. All tables were imported from both databases into statistical computing software and code was written to determine in which database the record was remediated and a combined remediation database was created. For combinations on variable bases, as corrections were only entered when the value was wrong, a function was created that took the original values of each variable and overwrote them only if a correction was made.

Electronic remediation of the data mostly involved variable simplification and cleaning. Looking at the frequency distributions of many of the variables in the original database during the electronic audit it became apparent that the database clearly lacked standardized rules for data entry. This resulted in many variables where the values were accurate but in the context of analyzing co-dependencies and building models, were far too heterogeneous and complicated to be useful. These variables ranged from variables needing simple normalization (making sure case use and spelling was consistent throughout the variable), to editing free text variables with dozens of values down to only a few useful values. When the data tables were exported from the original FilemakerPro database it appears a conversation error occurred where any number of "/v's" were inserted into variable values (before, after or in the middle of the value) and the cleaning effort was largely centered on eradicating them.

To make sure only data that was properly consented made it into the final database an algorithm was created that flagged all records that fell outside the window of consent. All records needed to first be linked to enable comparison to the cutoff date added to the Master Subject Index. **Figure 2** in **Appendix A** is a simple depiction of the relationships between the tables and lists which of the tables contain dates. The Patients and Family History tables both had no date but both were filled out using only information from the first visit so all records in these tables were fully consented (additional family history could have been added after the first visit, if these visits fell outside of the window of consent the Family History record was marked as "not in file" during manual remediation). The Visits and Exams tables both had date variables and although the Menstrual History and All Findings tables did not have date variables, dates for each record was easily derived from the Visits and Exams table respectively. The Surgeries table did contain a variable for date of surgery but not the date of the visit the surgery was recorded and visit IDs were not filled in consistently enough to link back to the visits table. It was therefore decided to include all surgery records where the date of surgery preceded the consent cutoff date. The Hormone History table proved to be somewhat problematic. The table was clearly set up originally were hormone history records should have been re-entered at every visit but in practice only the most up to date information (usually from the last visit) was entered in the database. Since it would have been very time consuming to enter all the missing records it was decided to keep this scheme with the caveat that if the use category changed from the last allowable visit to the last visit (e.g. from current to past) a new record was created and the old record was marked as "not in file" and given these considerations no Hormone History records were marked as un-consented.

Once combinations had been made, electronic remediation had been completed and all removal flags had been applied, a final version was extracted from each table. To be included in the final database a record could not have been marked as "not in file" during manual remediation or be flagged as falling outside of the consent window and had to be flagged as having been remediated. The entire preceding process was completed twice; once before manual remediation was fully completed to allow for the development of a modeling data set so that the modeling could be completed within the allotted time frame, and once when manual remediation was completed to create the fully remediated (Microsoft Access) database.

The basic multidimensional structure of the database meant that a flattened dataset with only one record per patient had to be derived to enable modeling. Based on insights from the original, un-remediated version of the breast screening model (what worked well, what could have been better), a list of 22 variables was devised to include in the initial modeling dataset. Since the end goal was to develop a model that used information determined at a women wellness exam; aside from the outcome, all variables in the modeling set should have been known at the time of the first visit. Variables that already had exactly one record per patient required no flattening and were kept "as is". To flatten bra cup size, breast density, palpable lesion, and menstrual cycle variables values were extracted from the first visit from corresponding variables in the database. Two flattened surgery variables were created: Prior Biopsy and Prior Non-Biopsy Surgery indicating whether or not the patient had had either a biopsy or a non-biopsy breast surgery prior to their first visit. To create this variable another variable had to first be created in the Surgeries table that compared the date of surgery to the date of first visit, and denoted whether the surgery took place before the first visit. Then a second variable was created that separated all surgery records into biopsy and non-biopsy. Using these two variables flat surgery variables were created (if a patient had any record in the Surgeries table that had a value of "Before" for the former variable and a value of "Biopsy" for the second their Prior Biopsy record would be "Yes" otherwise it would be "No"). Two different approaches were used to create a flattened family history of breast cancer variable. The first was a simple yes or no variable where a "Yes" meant that the patient had a record in the Family History table and a "No" meant they did not. The other was based on the closest degree relative with family history of breast cancer (first, second, higher or none). Since the original model included the use of hormones as an indicator of higher risk of developing breast cancer for all hormones represented in the Hormone History table, a yes or no variable was created indicating whether or not the patient had ever used that hormone. As the original data included no outcome variable, no variable that directly said this patient should or shouldn't get further breast screening so we needed to create one. It was decided that creating a variable based on BIRADS ratings would be the most accurate metric of whether or not the patient benefitted from the additional screening. The outcome variable, ShouldGetScreened, was developed by assigning any patients that had any All Finding records with a BIRADS rating of three or higher a "Yes", those whose highest BIRADS rating was less than three a "No" and those with no records in the All Findings table were left missing.

Training and Validation of Machine Learned Bayesian Belief Network

A Bayesian Belief Network (BBN) is a probabilistic directed, acyclic graphical model that represents a set of random variables and their conditional dependencies. Traditionally, BBNs are developed by domain experts based on their prior beliefs or prejudices. Automating this process using a proprietary machine learning software and thus removing as much bias as possible is the key to what was done. The ml-BBN are created using a heuristic algorithm learns from the data to quickly recognize a good and highly likely structure of the data's conditional dependencies.

The three-fold step-wise modeling process consisted of: preliminary modeling, global modeling and focused modeling. During preliminary modeling data quality was examined with patterns of missing data. Both overall quality and levels of missing data were studied extensively through the process of auditing and remediating and nothing new was revealed during this phase of the modeling. During global modeling identifying and removing any confounding, duplicate or unnecessary variables was sought. In our preliminary and global models some of the variables expected to be related to the outcome were not, most notably these included oral contraceptives and family history of breast cancer. It was at this point a simpler family history variable (yes/no) was created to see if the decrease in complexity would lead to a model where the two are connected. There have been studies that claim that family history of breast cancer is not a strong risk indicator for young women[1] a claim that the models seem to support. To make sure that the variables unconnected to the outcome were not caused by the influence of confounding variables numerous models were trained each time leaving out a variable or selection of variables to see if it had any effect on which variables were connected to the outcome. The result was always the same and so the final focused model as created. **Figure 3** in **Appendix A** shows the final model.

The variables that ended up in the final model aside from the outcome were age, breast density, palpable lesion, nulliparity, history (current or past) of Depo use, history (current or past) of hormone replacement therapy, prior biopsy, frequency of menstrual periods and duration of menstrual periods with the first three being first degree associates of the model.

To enable validation of the model, the modeling data was randomly segmented into a training set consisting of 80% of the data and a testing set consisting of the remaining 20%. After the model had been trained the testing set was recursively fed through the model and generated Receiver Operating Characteristic (ROC) curve and calculated the Area Under the Curve (AUC) as a metric of model performance. The AUC for recommended screening proved to be fairly robust at 0.735.

Prototype Web Deployment of Model

To facilitate possible future deployment of the new model a prototype web based decision support tool using the original (un-remediated) model was created. This online tool contains an interface to interact with the model enabling people to enter a patient's information and receive a risk estimate indicating whether the patient should get further breast screening.

Key Research Accomplishments

The EIS database was remediated and updated to fix data errors identified during a DCA audit, to account for updated consent statuses, and to develop a final Bayesian Relieve Network classification model to identify young women at higher risk of having or developing malignant or pre-malignant findings.

A web-based prototype of a Decision Support tool that can be used by patients and/or clinicians via a web browser was developed. This tool was designed to provide an intuitive graphical interface that will be served from a centrally hosted Decision Support server. The tool uses open-source platforms to the greatest extent possible, supports user log ins, personalized estimates of outcomes, and personalized risk assessments

Reportable Outcomes

- Study database audited to 1) remediate consent and input issues identified by DCI audit; 2) update study database to include additional visit records
- Issues identified in the study database remediated
- Data curation performed and a training set for classifier development completed
- Train and validate a machinet-learned Bayesian Belief Model to support a Decision Support tool
- Completed study cohort data set

Conclusion

The EIS database was remediated and updated to fix data errors identified during a DCI audit, account for updated consent statuses, and a final Bayesian Belief Network (BBN) classification model was developed to identify young women at higher risk of having or developing malignant or pre-malignant findings.

Appendix A- Figures

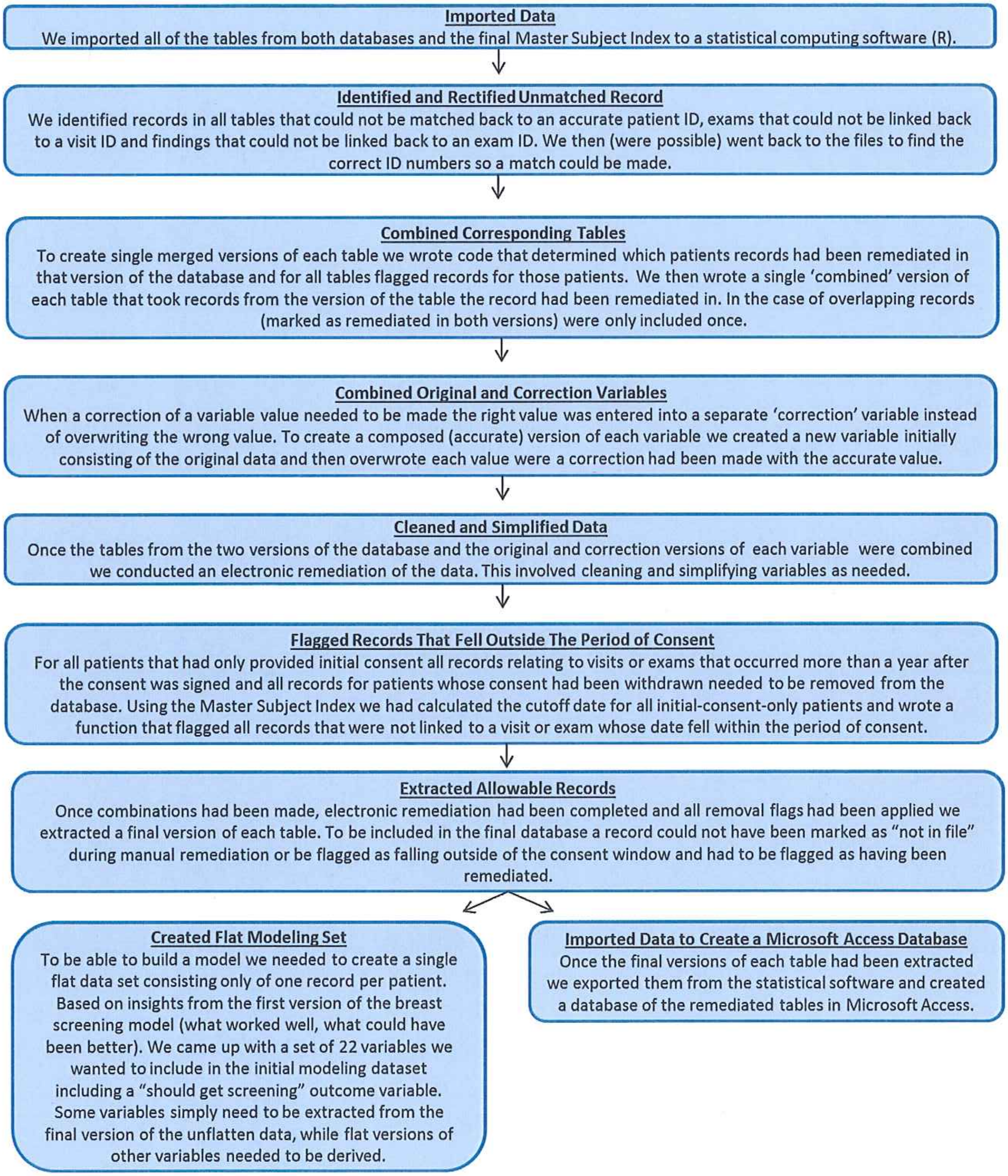

Figure 1 – Data Curation Flow Chart

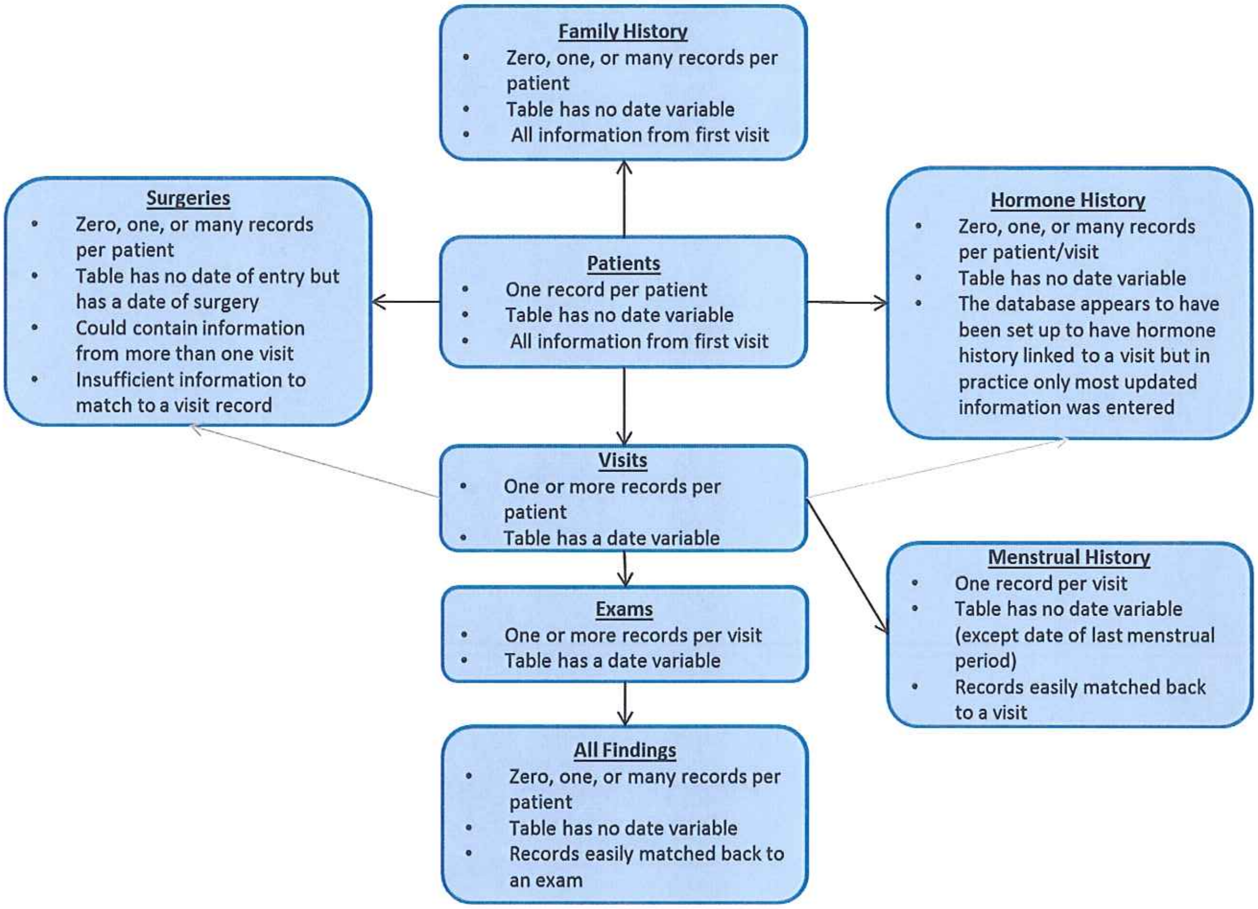

Figure 2 – Table Relationships

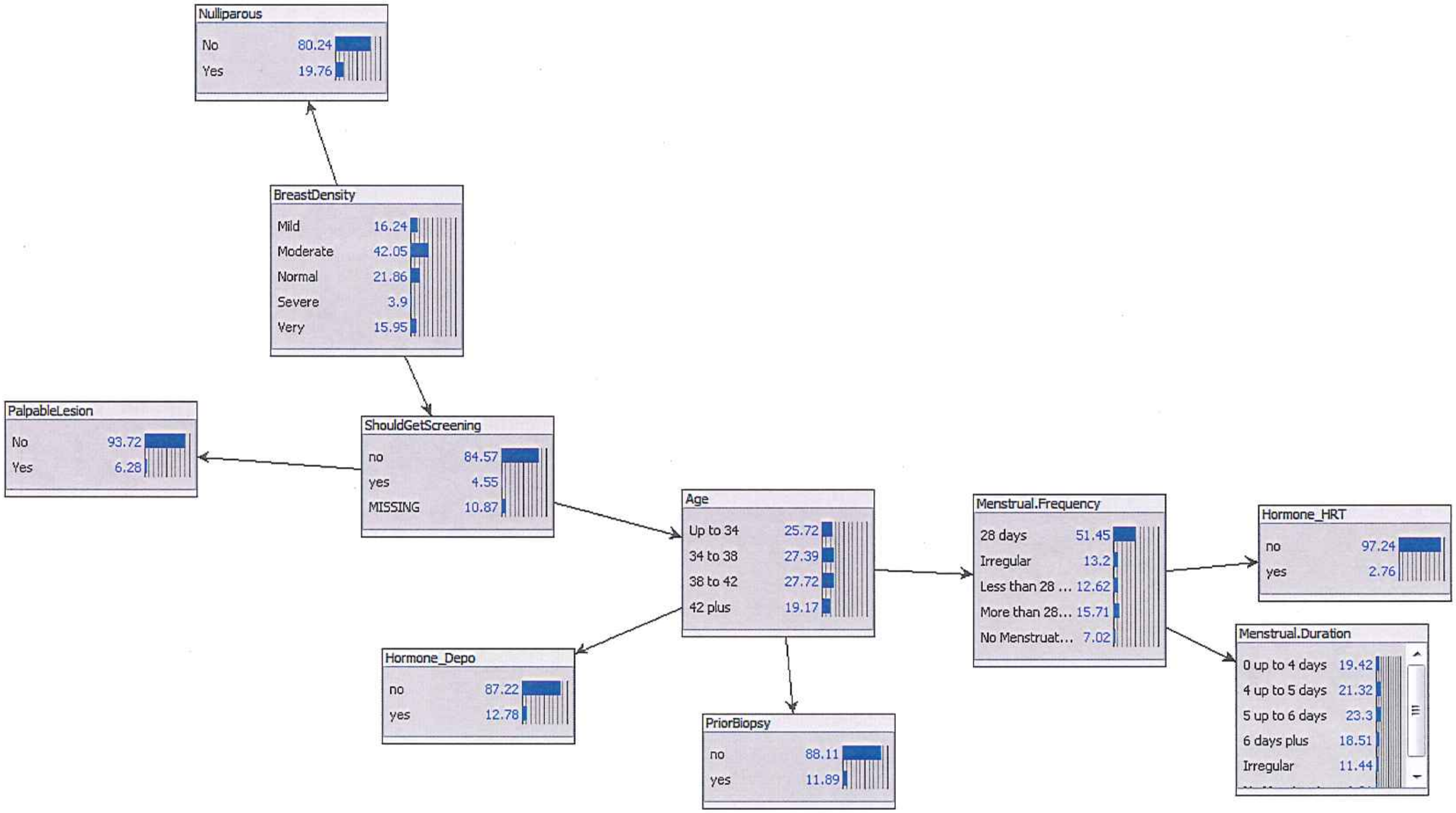

Figure 3 – Final Breast Screening ml-BBN

Appendix B – Data Remediation Process